Data Science in Action #2

Detecting Structural Changes and Outliers in Longitudinal Data

Gerhard Svolba
Data Scientist, SAS Austria

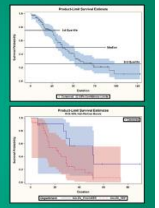
THE POWER TO KNOW.

Data Science Applications and Case Studies

Data Science
in Action: #1

Performing Headcount Survival Analysis for Employee Retention

Can assumptions about the average length of time intervals be made, even if most of the endpoints have not yet been observed?

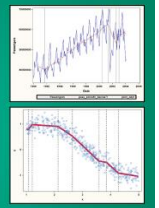


Survival analysis methods: Kaplan-Meier estimates
Cox Proportional Hazards regression
Survival Data Mining

Data Science
in Action: #2

Detecting Structural Changes and Outliers in Longitudinal Data

Can events and changes in the course over time be automatically detected?

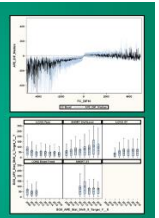


Smoothing Of Longitudinal Data
Multivariate Adaptive Regression Splines
Automatic Breakpoint Detection
Automatic Detection of Outliers with ARIMA Models

Data Science
in Action: #3

Explaining Forecast Errors and Deviations

Do the demand planners really improve forecast accuracy with their manual overwrites?

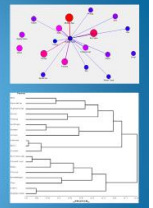


Linear Regression
Quantile Regression
Descriptive Statistics

Data Science
in Action: #4

Listening to Your Data – Discover Relationships with Unsupervised Analysis Methods

Can your data tell you stories about your analysis subjects, even if you don't ask explicitly?

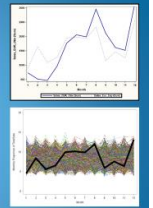


Unsupervised machine learning methods:
association analysis
variable clustering

Data Science
in Action: #5

Checking the Alignment with Predefined Pattern

Which customers show a behavior that is far from what you expected?

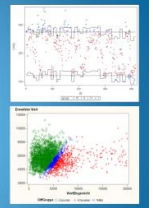


Chi2 independency test
Benford's law
Time Series Similarity

Data Science
in Action: #6

Proving a reference value that considers all available co-information

Can analytics help me to reduce the "Yes, but ... " sentences in my business discussions?

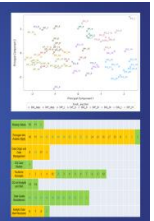


Linear Regression
Decision Trees
Time Series Analysis

Data Science
in Action: #7

Topic Search Documents and Clustering

Can I automatically find clusters of documents with similar content?

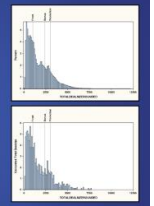


Text Mining
Text Parsing (Synonyme, Stemming, Stop-Listen)
Term by Document Weights

Data Science
in Action: #8

Using Monte Carlo Simulations to Understand the Outcome Distribution

When the sales manager looks at the project pipeline, does the sum of weighted averages give him or her a full picture?



Monte Carlo Simulations
Mathematical Programming

Data Science
in Action: #9

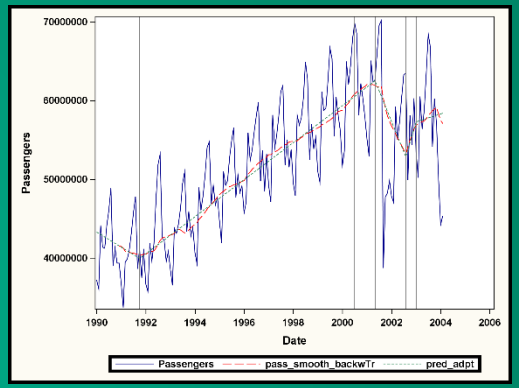
Studying Complex Systems – Simulating the Monopoly Board Game

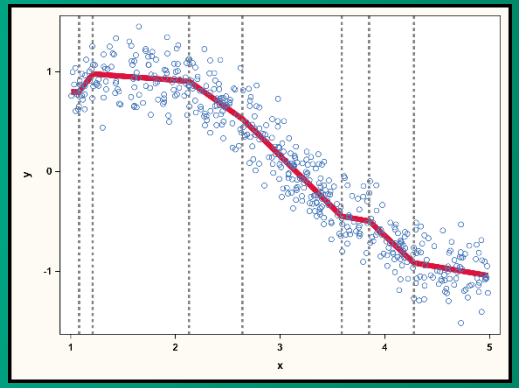
How can you simulate complex environments to get insight in the mos frequent processes?

Monte Carlo Simulations

Data Science in Action: #2

Detecting Structural Changes and Outliers in Longitudinal Data



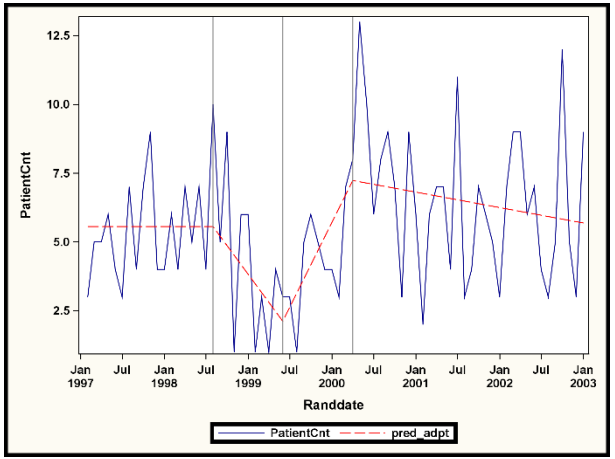


Can events and changes in the course over time be automatically detected?

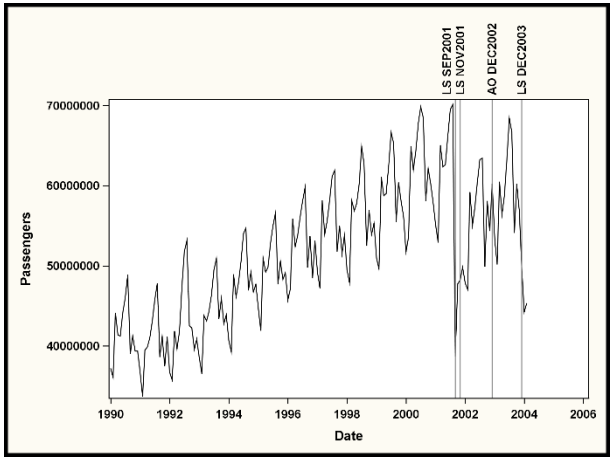
Smoothing Of Longitudinal Data
Multivariate Adaptive Regression Splines
Automatic Breakpoint Detection
Automatic Detection of Outliers with ARIMA Models

Automatically Detect Breakpoints and Outliers

Use machine learning methods to identify time points in your data where the course over time deviates from „normal“ behavior.



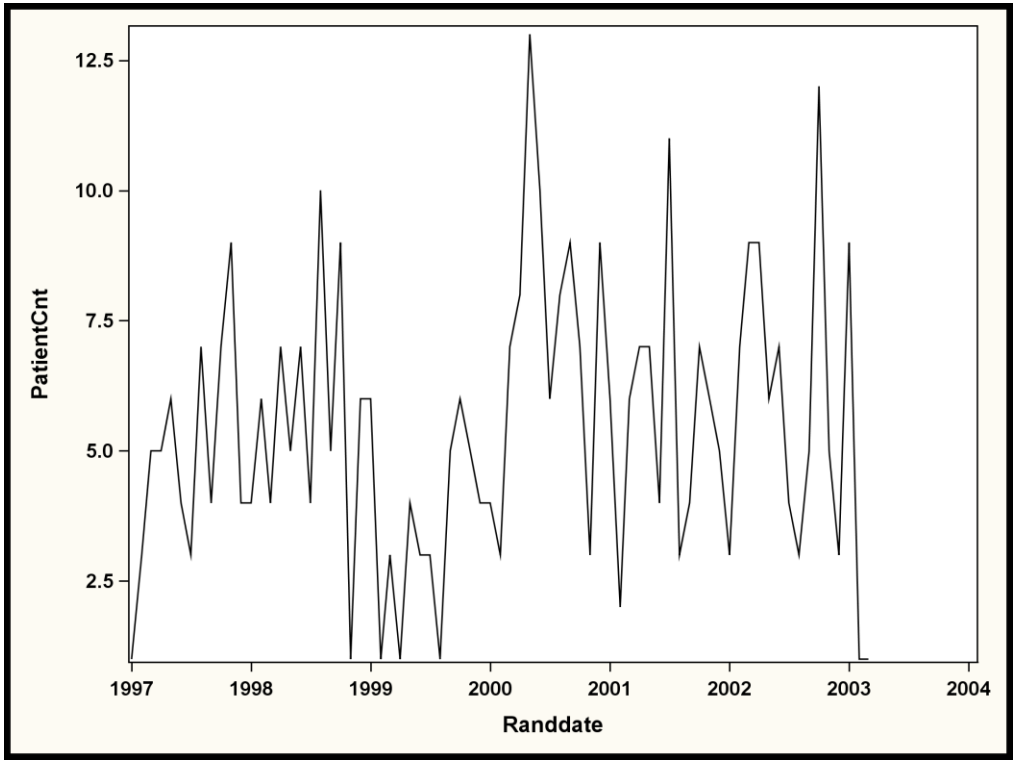
Use multivariate regression splines to identify breakpoints over time.



Detecting shifts and pulse events in your data with ARIMA Models.

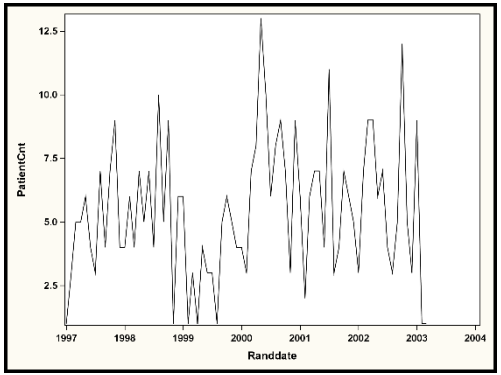
Detecting Breakpoints

Recruitment Numbers from a Clinical Trial

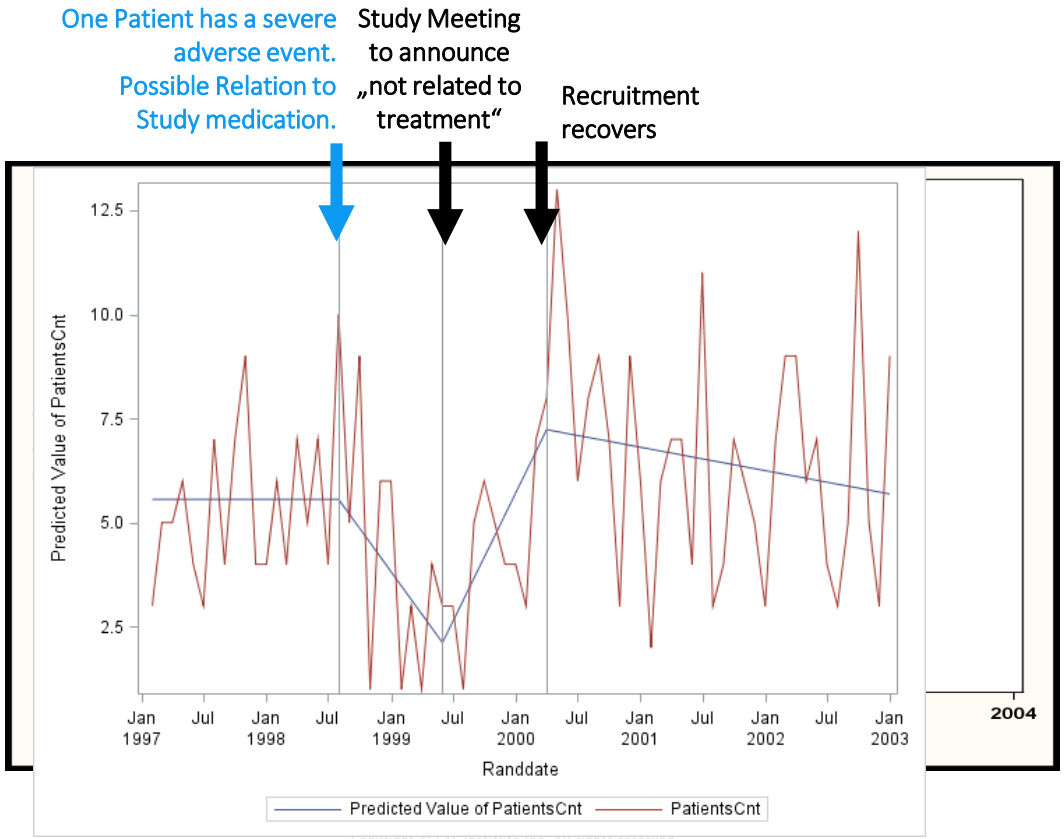


Can the breakpoints be detected automatically?

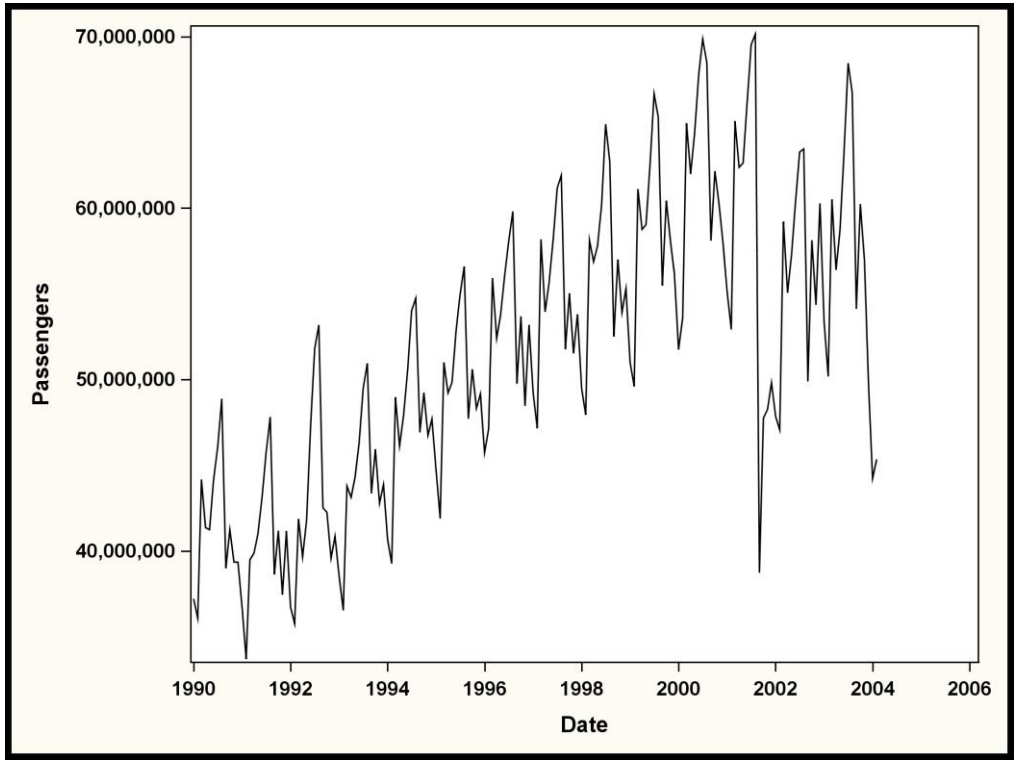
- Multivariate adaptive regression splines
  - Non-parametric regression techniques
  - Regression splines + model selection
  - ADAPTIVEREG procedure in SAS/STAT
- Breakpoints need not to be specified in advance.
- Learn from your data when a change might have taken place.



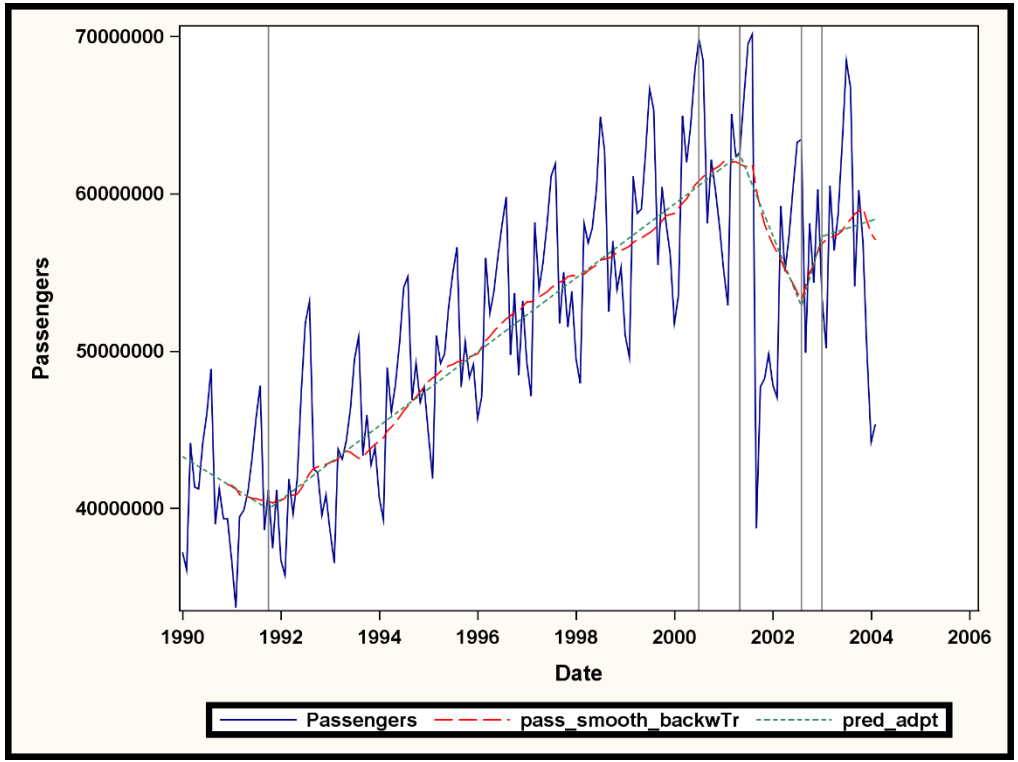
What happened in the clinical trial at certain points in time?



Airline Passenger Data (Monthly Sum)



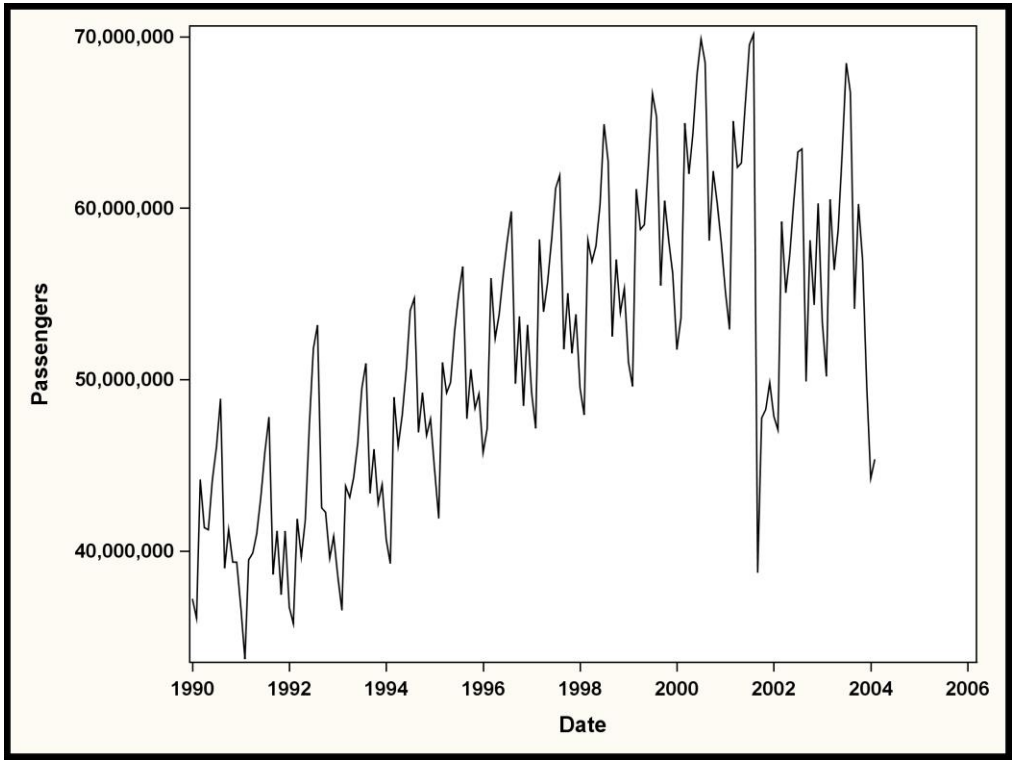
Adding the automatically detected knot points to the line chart



Knot
01OCT1991
01JUL2000
01MAY2001
01AUG2002
01JAN2003

Smoothing of Longitudinal Data

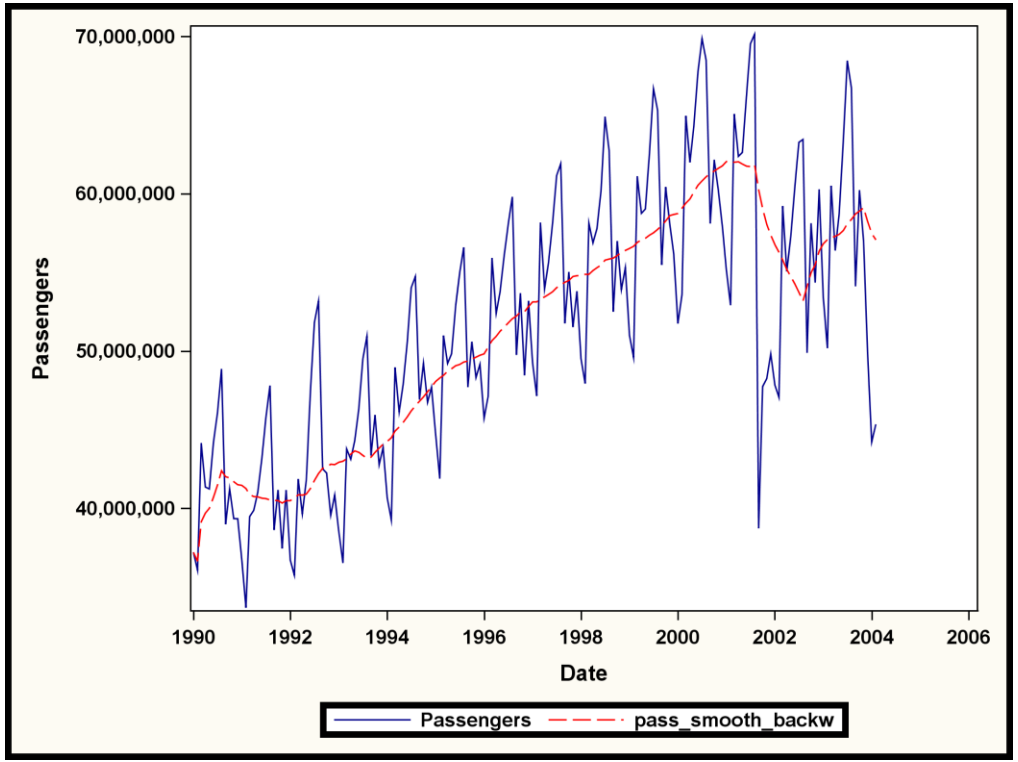
Airline Passenger Data (Monthly Sum)



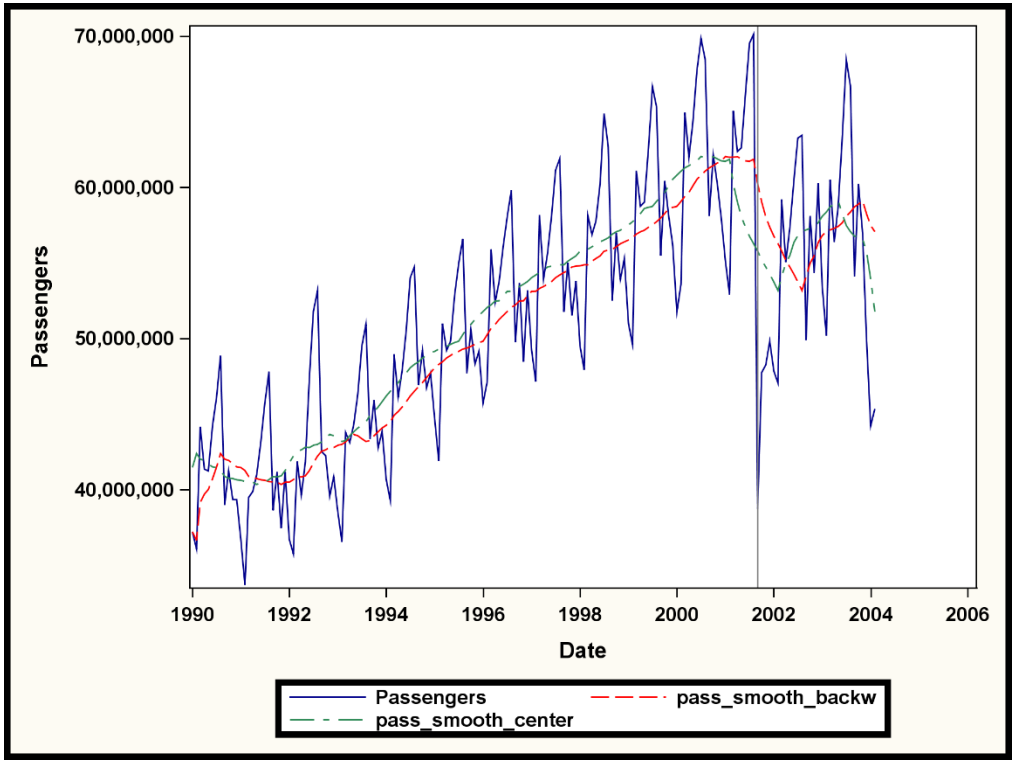
Centered and Backward Smoothing

Orientation Type	Feb 24th	Feb 25th	Feb 26th	Feb 27th	Feb 28th	Mar 1st	**Mar 2nd**	Mar 3rd	Mar 4th	Mar 5th
							Actual Date			
Centered Smoothing				-3	-2	-1	**0**	1	2	3
Backward Smoothing	-6	-5	-4	-3	-2	-1	**0**			

Backward Smoothing of the Airline Passenger Time Series



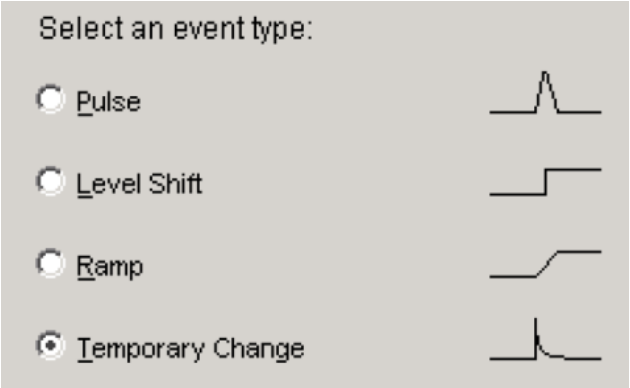
Backward and Centered Smoothing of the Airline Passenger Time Series



Detecting Outliers

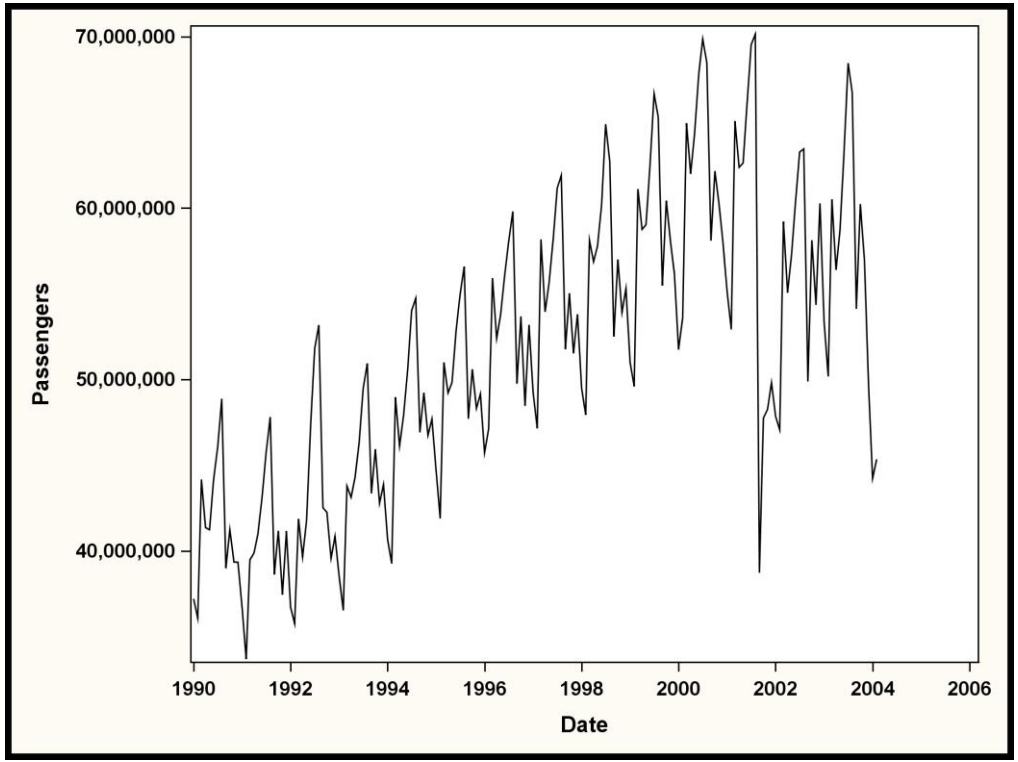
Procedure in Detecting Outliers

- Different types:
  - Outliers (pulse)
  - level shift
  - ramp
  - temporary change



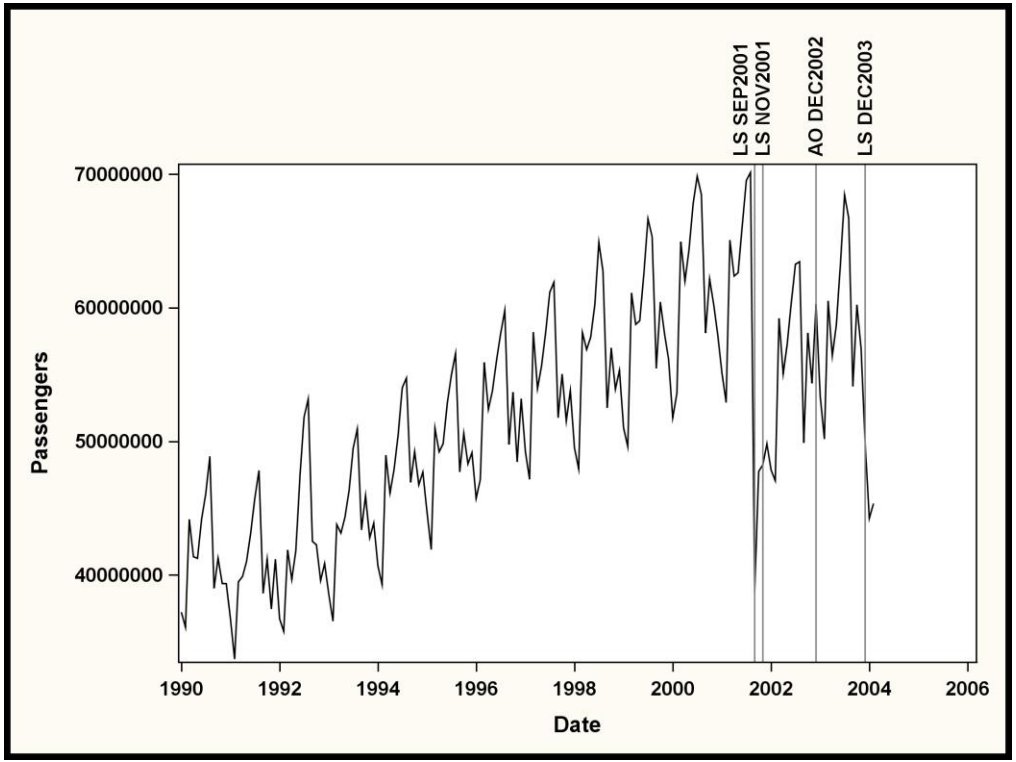
- Fit an ARIMA model to the time series with the X13, HPFDIAGNOSE or TSMODEL procedure
- Automatically identify those points where the time series deviates from the average pattern

Applying Outlier Detection to the Airline Passenger Data

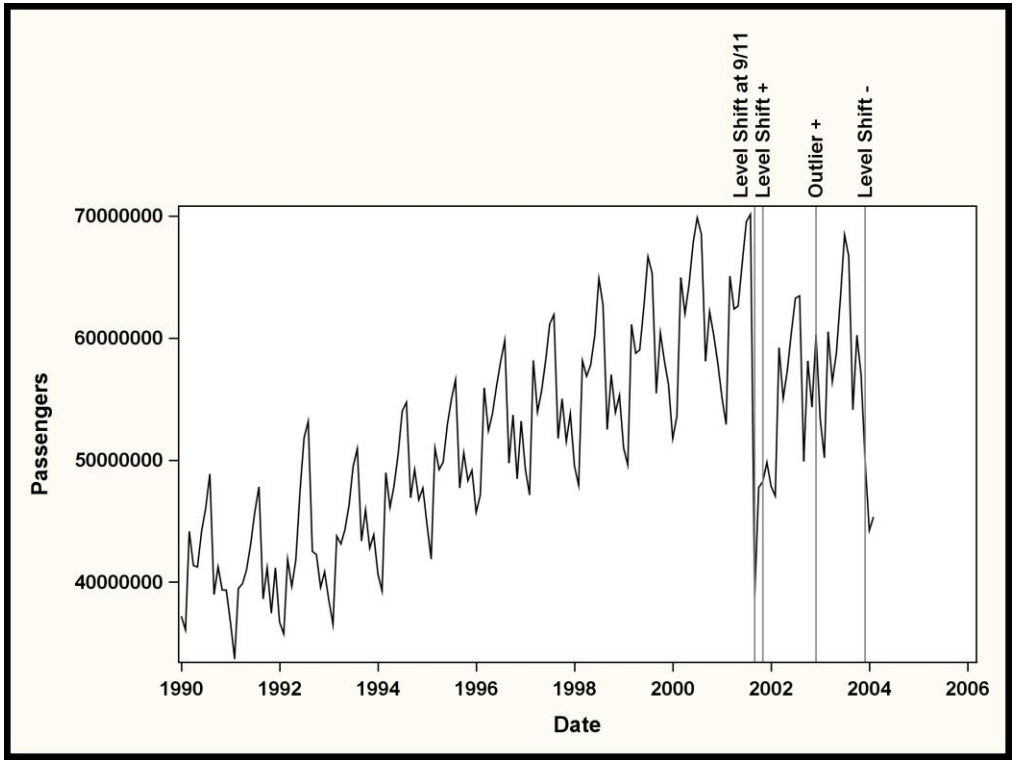


RegVar	Estimate
LS SEP2001	-17,993,817.7
LS NOV2001	5,939,640.5
AO DEC2002	5,039,786.8
LS DEC2003	-8,531,934.1

Labeling the Outliers in the Time Series Plot



Displaying individual labels

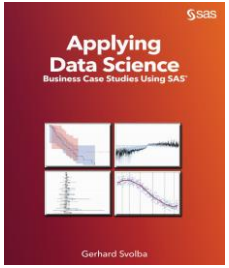


Conclusion

- Breakpoint and outlier detection can automatically detect structural changes in your time series data.
- Results are provided as lists or graphically
- Automatic insertion and labeling of the time points is key
- When smoothing time series data, be careful when selecting the smoothing method.

Analytics and Data Science is there to help you!

- Get a clearer, more objective picture of your data and your analysis subjects
- Get explicit results instead of searching the needle in the haystack
- Make your data talk to you!
- Receive findings automatically instead of manually
- Do it again! – treat models as an asset and repeat your analysis

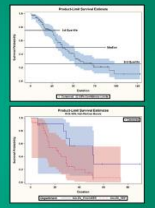


Data Science Applications and Case Studies

Data Science
in Action: #1

Performing Headcount Survival Analysis for Employee Retention

Can assumptions about the average length of time intervals be made, even if most of the endpoints have not yet been observed?

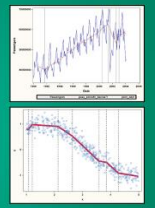


Survival analysis methods: Kaplan-Meier estimates
Cox Proportional Hazards regression
Survival Data Mining

Data Science
in Action: #2

Detecting Structural Changes and Outliers in Longitudinal Data

Can events and changes in the course over time be automatically detected?

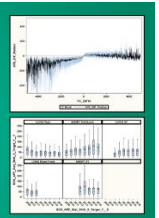


Smoothing Of Longitudinal Data
Multivariate Adaptive Regression Splines
Automatic Breakpoint Detection
Automatic Detection of Outliers with ARIMA Models

Data Science
in Action: #3

Explaining Forecast Errors and Deviations

Do the demand planners really improve forecast accuracy with their manual overwrites?

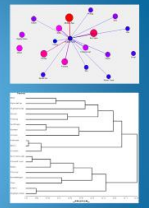


Linear Regression
Quantile Regression
Descriptive Statistics

Data Science
in Action: #4

Listening to Your Data – Discover Relationships with Unsupervised Analysis Methods

Can your data tell you stories about your analysis subjects, even if you don't ask explicitly?

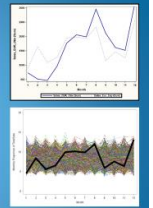


Unsupervised machine learning methods:
association analysis
variable clustering

Data Science
in Action: #5

Checking the Alignment with Predefined Pattern

Which customers show a behavior that is far from what you expected?

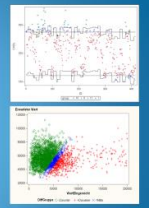


Chi2 independency test
Benford's law
Time Series Similarity

Data Science
in Action: #6

Proving a reference value that considers all available co-information

Can analytics help me to reduce the "Yes, but ... " sentences in my business discussions?

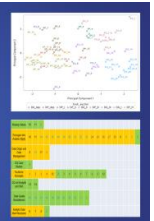


Linear Regression
Decision Trees
Time Series Analysis

Data Science
in Action: #7

Topic Search Documents and Clustering

Can I automatically find clusters of documents with similar content?

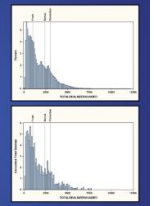


Text Mining
Text Parsing (Synonyme, Stemming, Stop-Listen)
Term by Document Weights

Data Science
in Action: #8

Using Monte Carlo Simulations to Understand the Outcome Distribution

When the sales manager looks at the project pipeline, does the sum of weighted averages give him or her a full picture?



Monte Carlo Simulations
Mathematical Programming

Data Science
in Action: #9

Studying Complex Systems – Simulating the Monopoly Board Game

How can you simulate complex environments to get insight in the mos frequent processes?

Monte Carlo Simulations

Get access to more content:

SAS DACH @Youtube: https://www.youtube.com/user/SASsoftwareGermany

Blogs on LinkedIn: https://www.linkedin.com/in/gerhardsvolba/

Twitter: https://twitter.com/gsvolba

Content on Github: https://github.com/gerhard1050

Books @SAS-Press: https://support.sas.com/svolba